DOCUMENT RESUME

ED 186 444 TM 800 081

AUTHOR Goldberg, Jill; Hulin, Charles L.
TITLE An Empirical Study of the Use of Item-Item Curves to Detect Item Bias: Preliminary Report.
PUB DATE 21 Sep 79
NOTE 41p.; Paper presented at the Annual Meeting of the American Psychological Association (87th, New York, NY, September 1-5, 1979).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *Item Analysis; Latent Trait Theory; Part Time Employment; Questionnaires; Sex Differences; *Test Bias
IDENTIFIERS *Item Item Curves; Job Descriptive Index

ABSTRACT

A Job Descriptive Index was studied for possible bias by examining item-item curves for various pairs of items computed for contrasting samples. The two pairs of contrasting samples were 669 part-time workers versus 920 full-time workers and 486 male workers versus 1,146 female workers in a large merchandising company. Item-item curves were computed by obtaining the scores for all the workers on the 18-item, 3-choice scale; then stratifying each sample on the basis of total score into groups of about 50 each; finding the percent of yes responses in each group for each item; and finally plotting percents of yes responses on various pairs of items for the two contrasting samples on the same pair of item axes. In general, the items had closely similar curves for the contrasting samples except for one item that was chosen much more frequently by part-time workers than by full-time workers. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED186444

An Empirical Study of the Use of Item-Item Curves to Detect Item Bias: Preliminary Report

Jill Goldberg

Charles L. Hulin

University of Illinois

at Urbana-Champaign

September 21, 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jill Ellen Goldberg

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: Item-Item Curves

TM800 081

Abstract

This study is a preliminary empirical investigation of the usefulness of Item-Item analysis to detect item bias. Item bias exists when two examinees, with equal ability (θ), have unequal chances of getting an item correct depending on their group membership. Item-Item analysis provides a simpler approach to detect item bias than has been used in the past.

Latent Trait Theory is a measurement theory that analyses the responses of individual examinees to individual items. Analysis of the response vectors of the examinees, yields the ability level, Θ, of each examinee and the necessary item parameters. Latent Trait Theory makes two assumptions, unidimensionality of the latent trait space and local independence.

The assumption of unidimsionality simply means that a test or scale is measuring a single trait. A unidimensional latent trait space must be assumed in order to apply the methods of latent trait theory.

The assumption of local independence states that the response an examinee gives to one item is independent of the response he gives to any other items on the exam; the joint probability of an item response vector is equal to the product of the probabilities of the responses to each item.

From the assumptions of unidimensionality and local independence the a, b, and c item parameters, for the three parameter logistic model, should be invariant across groups or samples. Therefore, in principle, it should not matter on what group item parameters are computed. Whatever the sample, parameters should be the same, except for estimation error, if the parameters are in the same metric.

If the a or b or c parameter values are significantly different when determined separately in two groups and converted to the same metric, it means that examinees with the same estimated abilities (Θ's) will have different chances of getting the item correct depending on their group membership. Such a situation is clearly unfair, and is offered as a definition of item bias. Thus, item bias between groups A and B for item i can be

defined as $P(i_A|\theta=k) \neq P(i_B|\theta=k)$ where k is some value of θ. That is, the probability of a correct response to the ith item, given $\theta=k$ for group A is not equal to the probability of a correct response to the same item given $\theta=k$ for group B.

The probabilities of getting an item correct need not be unequal at all θ values for item bias to exist. For instance, an item may be biased at high values of θ and not biased at low values of θ. This is because the probability of getting an item correct may be different due to any one of the three item parameters being different.

If item bias does occur, it is desirable to know how a particular item is biased and what effect that bias has on each group of examinees. Lord (1977) discusses one method used to detect item bias as the comparison of Item Characteristic Curves (ICC's) of the item for each group.

An ICC is a mathematical function used in Latent Trait Theory which relates the probability of a correct response on an item to the trait measured by the item. An ICC uses the item parameters to determine the probability of a correct response. More specifically, an ICC plots the probability of getting an item correct on the vertical axis as a function of ability (θ).

By plotting the ICCs for two samples on the same graph, it is easy to detect the places where one curve deviates from another. If the curves of the two samples deviate from one another, over a given range of θ, then the item is biased over that range of θ for those two samples. An illustration of this, shown in figure 1, compares the ICCs of an actual item from an experimental form of the Coast Guard Selection Test, calibrated separately for blacks and whites (Warm, 1978). This figure shows that the probability of getting this item correct is greater for blacks than for whites at most values of θ.

- - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - -

It is important to note, that to use ICCs to detect item bias, the item parameters and ability for each examinee must be estimated. The estimation of these parameters is not a trivial problem.

Therefore, this paper proposes an alternative and simpler approach to detecting item bias. This procedure is called Item-Item analysis. Item-Item analysis is a new application of Latent Trait Theory proposed by Levine (1976). It involves obtaining an Item-Item Curve (I-I curve) as opposed to an ICC, both shown in figure 2. An I-I curve shows the probability of a correct response to one item against the probability of a correct response to another item, conditioning on the same value of θ.

- - - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - - -

Item-Item curves tell us whether or not an item is biased in one of two samples in the same way the ICC does. When two I-I curves are plotted for two samples of examinees, and the curves deviate from one another, beyond the limits of sampling, then item bias exists in at least one sample over some range of θ.

Item-Item curves are simpler to use because they do not require item parameters or examinees' abilities be estimated. This is possible because I-I curves are not plotted directly against θ as ICCs are, but involve the plotting of two probabilities, <u>both conditioned on the same θ</u>. The exact θ for each examinee need not be known because the double conditioning, in

effect, holds θ constant and the points on the I-I curve do not refer to any specific θ values. Errors in estimating θ at some value in one group are assumed to be equal to the error in estimating the same θ in another group.

In contrast, an ICC is plotted against θ, and therefore, the points on the ICC do refer to specific θ values; these θs must be estimated. To estimate θ as well as the item parameters, a computer program, LØGIST, was developed by Wood and Lord (1976). However, LØGIST requires the assumptions of latent trait theory, as well as the availability of high speed, and high capacity computer facilities. In addition, for convergence to occur, Lord recommends a reasonably large number of items, say 40 or more, and a reasonably large sample, say 1000 or more.

Because I-I curves do not depend on an exact estimation of θ, a formula score can be used to approximate θ for each examinee in order to plot the points of the I-I curves. Formula scores generally refer to a score obtained from any specified scoring formula other then the simple zero-one scoring scheme. Lord and Novick (1968) note that it is common practice to use a formula score as an approximation of θ. Using the formula score as an approximation of θ, the points of the I-I curve are obtained by computing the probability of a correct response given each estimated θ. Note, however, that the use of a formula score to estimate θ for ICC's becomes problematical. Since there is no one-to-one relationship between formula scores and θ, formula scores still must be converted to θ in order to obtain the ICC because it is necessary to know the points on the θ axis.

Note that with two ICC's showing the responses to two items administered to one population, we can obtain an I-I curve by reading off and plotting the probabilities of correct responses to the two items given the same level of θ

for each point on the I-I curve. One can derive I-I curves from two ICCs but not the reverse because the derivation of ICC's requires that item and individual parameters be estimated.

Using I-I curves to examine bias in a particular item is done by comparing the I-I curves from two groups, using common items for the horizontal and vertical axes of each curve. When I-I curves are generated for a group of items, one of the items serves as the common horizontal axis for all curves. The item chosen as common to all I-I curves should be one that is expected to have variance in examinee responses and one that seems to have the capacity to discriminate between examinees with different levels of the trait being measured. The discriminatory power of this item can be examined from the responses of examinees in the sample under investigation. Note that compositor of items could be used rather than single items. One item was used in this investigation for simplicity.

If two I-I curves for two different samples (same items) deviate from one another, there is evidence that item bias exists for a given group, most probably in the item on the vertical axis (since the horizontal axis of all I-I curves is the same item).

To be certain which item is biased, however, both items from the I-I curve showing bias must be plotted against another common item. For instance, suppose items i (horizontal axis) and j (vertical axis) are plotted against one another for samples A and B and there appears to be a bias. Item i (vertical axis) should be plotted against item k (horizontal axis) and item j (vertical axis) can also be plotted against item k (horizontal axis) to determine whether item i and item j is the biased item. This is shown in figure 3. If no bias occurs when item i is plotted against item k, but bias

does occur when item j is plotted with item k, then this is evidence that item j is the biased item.

- - - - - - - - - - - - - - -

Insert Figure 3 about here

- - - - - - - - - - - - - - -

Subjects

Questionnaires containing the Job Descriptive Index (JDI), a job satisfaction inventory developed by Smith, Kendall, and Hulin (1969), were administered nationally in the summer and fall of 1977 to 1,632 employees of a large merchandising company.

Procedure

In our empirical investigation of the use of Item-Item analysis to detect item bias, we used seven items from the work scale of the JDI, shown in figure 4. The items analyzed thus far in this study were fascinating, routine, satisfying, respected, hot, healthful, and challenging.

- - - - - - - - - - - - - - -

Insert Figure 4 about here

- - - - - - - - - - - - - - -

The data collected from the employees was scored for each of the five scales of the JDI as follows:

3 for a positive response; 1 for uncertain; and 0 for a negative response.

The overall score for each scale was used to represent the formula score and to approximate θ.

The subjects were successively divided into sub-samples. The sub-samples already examined in this investigation of item bias were part time

vs. full time employees (669 part time workers and 920 full time workers), and male vs. female employees (486 male workers and 1146 female workers). Other sub-samples that are being studied will be based on race and education levels. Items from the other four scales of the JDI will also be examined for item bias within each of the sub-samples.

Once the subjects were divided into sub-samples, a frequency distribution was obtained of the number of examinees with each total work scale score. The individuals of each sample were then divided into groups, on the basis of formula scores, of approximately 50 examinees according to their total work scale score. The probability of a positive response to each item was then computed conditioning on each formula score level, to obtain the points for different I-I curves.

Of the items on the work scale of the JDI, the item "satisfying" was the most direct question concerning job satisfaction, and yielded the most variance in responses and discriminated best among the employees sampled. For this reason, the I-I curves were plotted with "satisfying" as the horizontal axis of all the I-I curves generated.

In addition, one set of item parameters were derived from LØGIST based on a large, heterogeneous sample (i.e. not a sample divided into part and full time employees). These parameters were used to derive points for theoretical I-I curves, again with the work scale item "satisfying" as the common horizontal axis. Theoretical I-I curves were derived to verify the shape and orientation of the empirical I-I curves as well as to reaffirm the existence or nonexistence of item bias. If item bias existed, the theoretical I-I curve was expected to fall between the I-I curves from the two sub-

samples. If item bias was nonexistent, the theoretical I-I curve was expected to fall on the I-I curves from the two sub-samples.

In addition to repeating the above, using JDI scales other than the work scale, as well as using other sub-samples, other procedures are being used. By using various sub-samples (part time vs. full time, black vs. white) in the Item-Item analysis study, some items will be biased for some sub-samples, while non-biased for other sub-samples. Taking the groups, for which bias exists for certain items, other items that appear not to be biased for that group, will be studied to determine the content of items that yield biased I-I curves for each sub-sample.

We will consider also another procedure that will challenge the sensitivity and efficiency of Item-Item Analysis to detect small differences. This procedure involves making mistakes purposefully. That is, as shown in figure 5, item 2 will be plotted against item 1 for sample A, but item 3 (instead of item 2) will be plotted against item 1 for sample B. If items 2 and 3 are very different from one another in content and parameters, such as the two work scale items "simple" and "necessary," then the two I-I curves for samples A and B should deviate from one another, (since two different items were used) indicating the apparent existence of item bias. However, a truer test of sensitivity will be when item 4 is plotted against item 1 for sample A and item 5 (instead of item 4) is plotted against item 1 for sample B, and items 4 and 5 are close in meaning, such as the work scale items "satisfying" and "pleasant". The two I-I curves from the two samples should still deviate from one another since different items were used. Item-Item Analysis should be sensitive enough to show this difference by yielding

two I-I curves that deviate from one another for samples A and B. The deviations should be smaller the more similar in meaning and intensity the two items are.

- - - - - - - - - - - - - - -

Insert Figure 5 about here

- - - - - - - - - - - - - - -

Finally, another possible procedure to confirm that Item-Item Analysis does, in fact, detect item bias between two sub-samples will be to use LØGIST to estimate the item parameters of each item for each sub-sample. If Item-Item Analysis yields evidence that bias exists for a certain item between two sub-samples, then LØGIST should yield different item parameters for that item when the parameters are estimated within each of the sub-samples. This would confirm the findings that support the Item-Item Analysis approach to detecting item bias.

Results

Most items appeared unbiased for the part time as compared to the full time samples of examinees. That is, the curves from the two samples had the same shape and orientation. For instance, figure 6 shows the results of the item "respected" plotted against the item "satisfying." The points from both the part time and full time samples fall along the same line, indicating that no bias exists for this item between the part and full time sub-samples.

The same is true of the I-ī curve when "routine" is plotted against "satisfying," shown in figure 7. Again, the points from both the part time and full time samples fall along the same line, indicating that no bias exists for the item "routine" between the part time and full time workers.

- - - - - - - - - - - - - - - - - -

Insert Figures 6 and 7 about here

- - - - - - - - - - - - - - - - - -

However, when the item "hot" was plotted against "satisfying" for the part time and full time samples, "hot" appeared to be a biased item, especially at the lower satisfaction levels. Figure 8 shows this bias as the deviation of the I-I curves for each sub-sample. The two I-I curves deviate from one another beyond the limit of sampling.

- - - - - - - - - - - - - - - -

Insert Figure 8 about here

- - - - - - - - - - - - - - - -

In order to be certain that the item "hot" was biased and that the item "satisfying" was not, both the items "hot" and "satisfying" were plotted against a third item, "respected." When the item "hot" was plotted against "respected," shown in figure 9, the bias between the part time and full time samples previously suggested was confirmed. The item "hot" still appeared to be biased at the lower satisfaction levels.

- - - - - - - - - - - - - - - -

Insert Figure 9 about here

- - - - - - - - - - - - - - - -

On the other hand, when "satisfying" was plotted against the item "respected", shown in figure 10, there appeared to be no bias in either the part time or full time samples. There was no deviation of the two I-I curves of each sub-sample for one another.

- - - - - - - - - - - - - - -

Insert Figure 10 about here

- - - - - - - - - - - - - - -

The theoretical I-I curves were used to confirm further the results that the item "hot" was biased for the part time and full time samples. Figure 11 shows that not only did the two I-I curves from the part time and full time samples, plotting "hot" against "satisfying," deviate from one another, but the theoretical I-I curve fell between the I-I curves for the part time sample and the I-I curve for the full time sample. Since the theoretical I-I curve was computed from parameters obtained from the entire sample, the theoretical I-I curve should fall somewhere between the I-I curves of the two sub-samples, part time vs. full time employees, when bias exists. Warm (1978) in his discussion of item bias, mentioned that the ICC of the "combined" group generally falls between the ICCs of the separate groups if bias exists.

- - - - - - - - - - - - - - -

Insert Figure 11 about here

- - - - - - - - - - - - - - -

For the unbiased items, the theoretical I-I curve fell on the I-I curves from each sample. For instance, figure 12 shows the plot of the item "healthful," determined to be unbiased. The I-I curves from both the part time and full time samples match each other as well as the theoretical curve.

- - - - - - - - - - - - - - -

Insert Figure 12 about here

- - - - - - - - - - - - - - -

For the male and female samples, none of the seven selected items of the work scale showed any bias between these two samples. For instance, figure 13 shows the plot of "hot" against "satisfying" for the male and female samples. The I-I curves for these two samples had the same shape and orientation. Another example is the plot of "challenging" against "satisfying", shown in figure 14, which also shows curves approximating one another for the male and female samples.

- - - - - - - - - - - - - - - - - - -

Insert Figures 13 and 14 about here

- - - - - - - - - - - - - - - - - - -

Discussion

In summary, when Item-Item Analysis was applied to data generated by the JDI administered to a large sample of workers, evidence was found suggesting that certain of the items were biased. It is not surprising that bias was detected in some of the work scale items. What is surprising, is that more evidence of bias was not found considering the restricted samples used in the construction of the JDI, its age, and the heterogeneity of the present samples available for Item-Item Analysis. One explanation for the little evidence of bias is that possibly the JDI is not susceptible to item bias as defined in this investigation.

Another possible explanation is that there is not enough difference between the part time-full time sub-samples or between the male-female sub-samples to yield bias. The sub-samples of race and education level, currently being studied, may prove to be more diverse and therefore yield more evidence of item bias.

Another major concern with this investigation is the sensitivity and efficiency of Item-Item Analysis to detect small differences. This will be the purpose of using different, and "wrong," items to derive I-I curves.

The procedure of Item-Item Analysis is relatively easy. It conserves scarce resources in terms of computer time. Item parameters do not have to be estimated nor are we concerned with precise estimates of ability levels of individuals. These estimation procedures require both extensive time and computer capacity. They also require the assumptions of latent trait theory regarding unidimensionality, local independence, and so on. In the case of the total JDI we know these assumptions to be suspect. They are not suspect for each scale considered separately, but the number of items in each scale (9 or 18) is too small to allow accurate parameter estimation using LØGIST. However, the questions of true efficiency and sensitivity of the procedure to small differences remain. We know the procedure has a considerable amount of promise. We do not yet know its efficiency.

Levine, M. V. Item-Item curves and consistent mental test parameter estimates. *Research Bulletin 76-36*, Princeton, N. J., Educational Testing Service, 1976.

Lord, F. M. Practical applications of Item Characteristic Curve Theory. *Journal of Educational Measurement*, 1977, 14, 117-138.

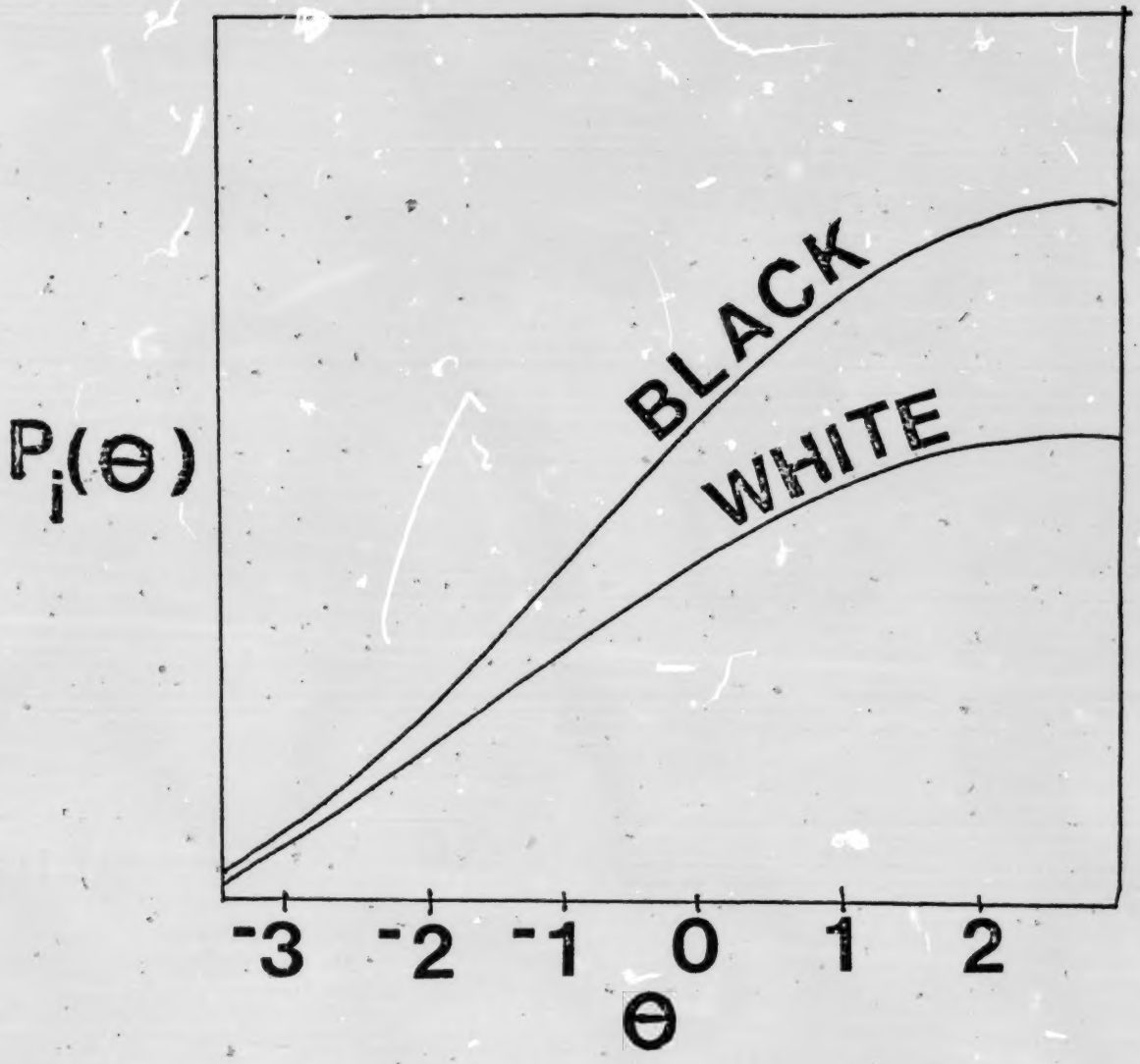
Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, Mass., 1968.

Smith, P., Kendall, L. M. and Hulin, C. L. *The Measurement of Satisfaction in Work and Retirement*. Rand McNally, Chicago, 1969.

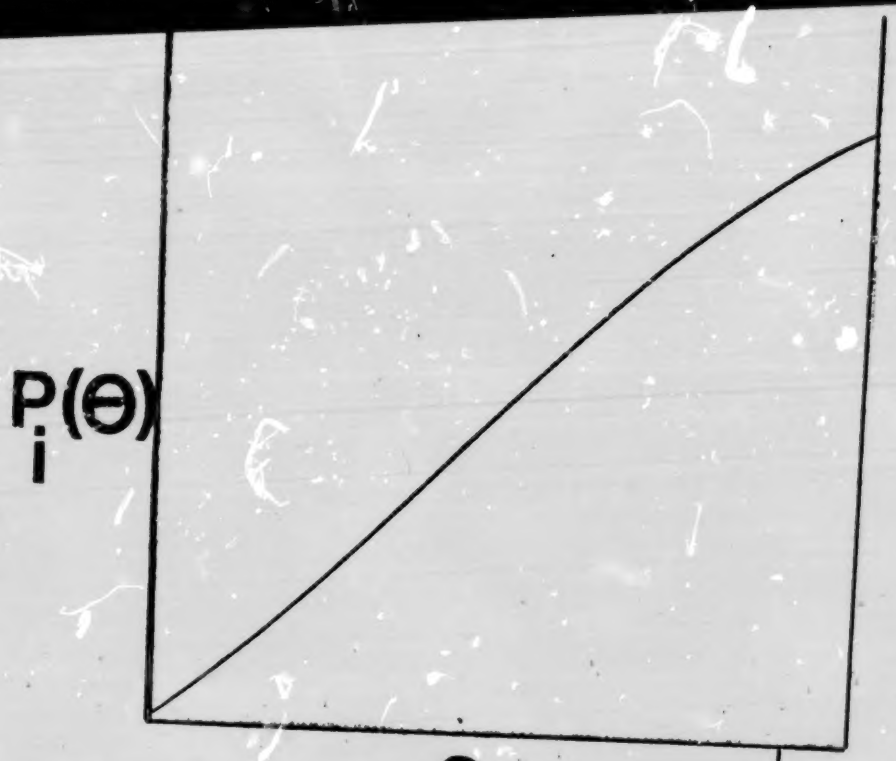
Warm, T. A. *A Primer of Item Response Theory*. U. S. Coast Guard Institute, Oklahoma City, Oklahoma, Technical Report 94, October, 1978.

Wood, R. L. and Lord, F. M. A user's guide to LOGIST. *Research Memorandum 76-4*. Princton, N. J., Educational Testing Service, 1976.

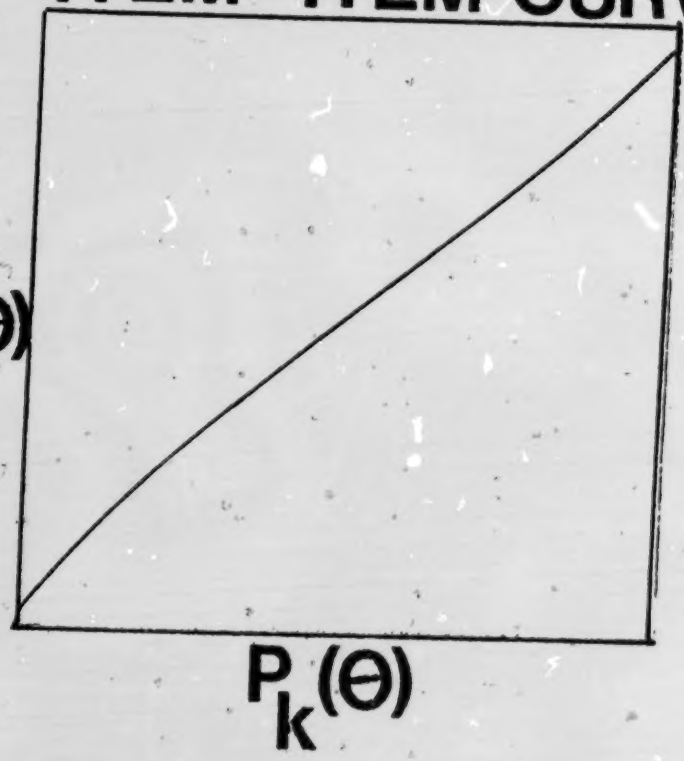
FIGURE 1



ITEM CHARACTERISTIC CURVES



ITEM CHARACTERISTIC CURVE



ITEM-ITEM CURVE

FIGURE 2

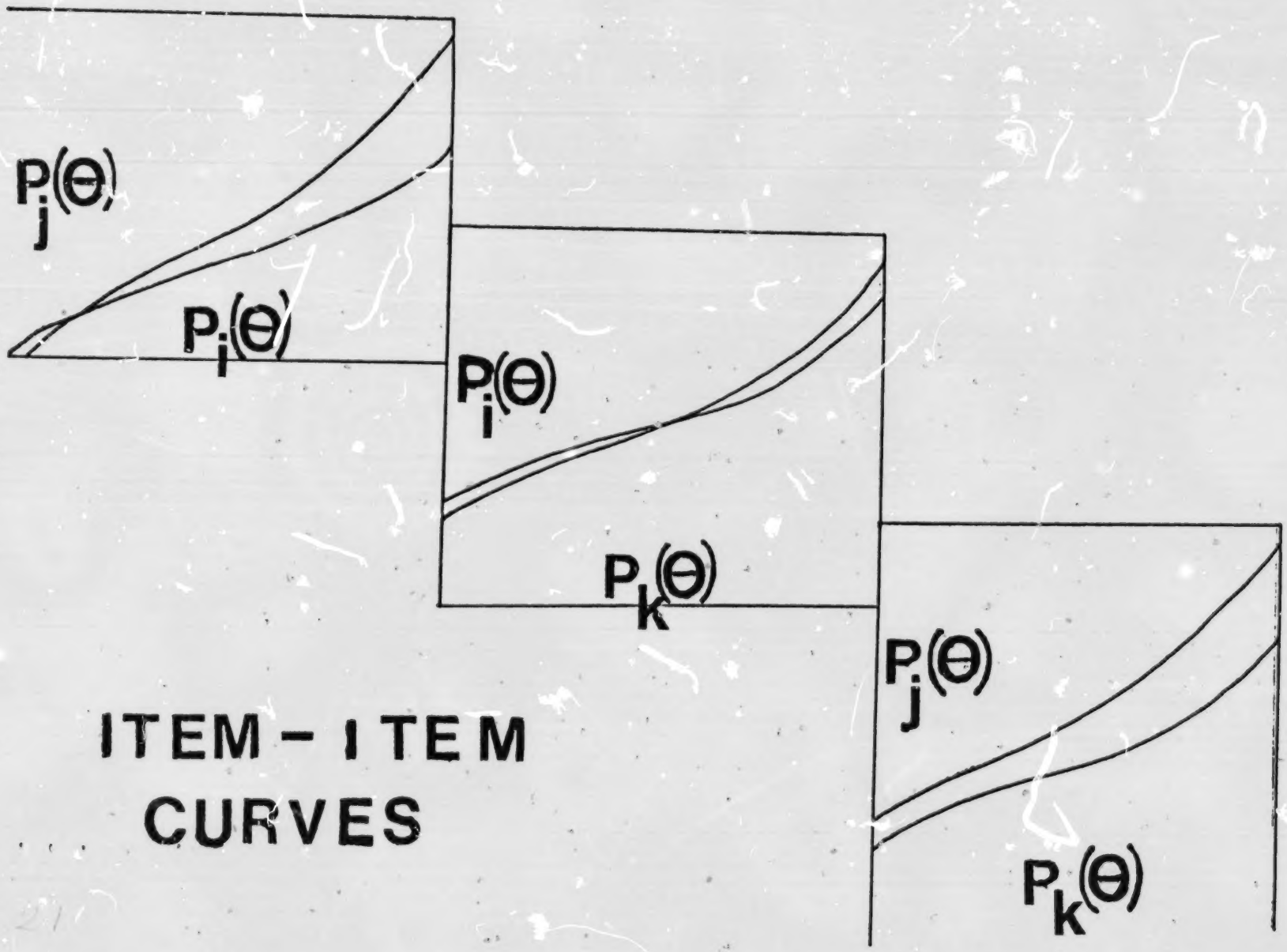


FIGURE 3

ITEM - ITEM CURVES

FIGURE 4

JOB DESCRIPTIVE INDEX

Y FOR "YES" IF IT DESCRIBES IT

N FOR "NO" IF IT DOES NOT DESCRIBE IT

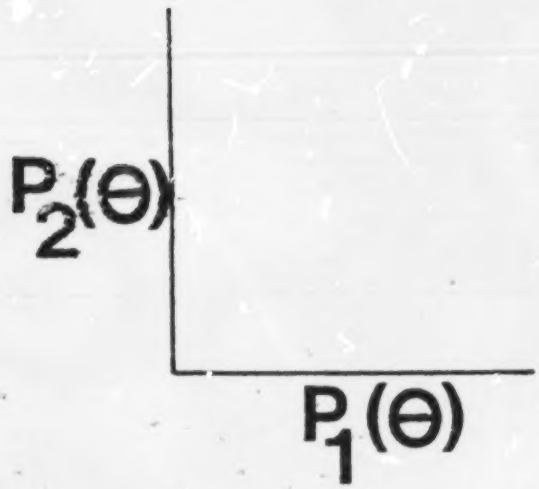
? IF YOU CANNOT DECIDE

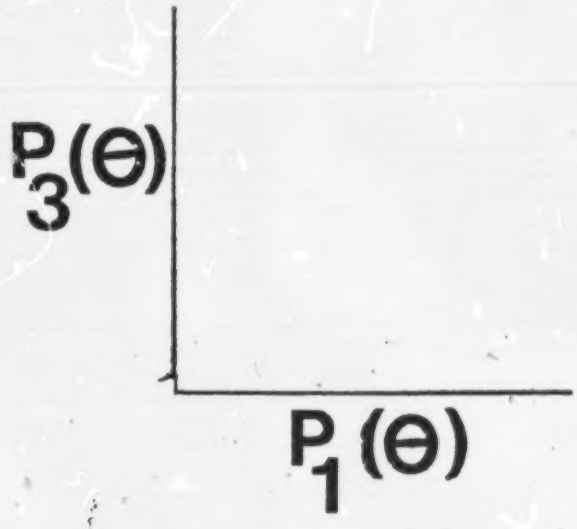
1. THE WORK I DO

___FASCINATING	___RESPECTED	___CHALLENGING
___ROUTINE	___HOT	___ON YOUR FEET
___SATISFYING	___PLEASANT	___FRUSTRATING
___BORING	___USEFUL	___SIMPLE
___GOOD	___TIRESOME	___ENDLESS
___CREATIVE	___HEALTHFUL	___GIVES SENSE OF ACCOMPLISHMENT

FIGURE 5

I. ITEMS 2 & 3 DIFFERENT

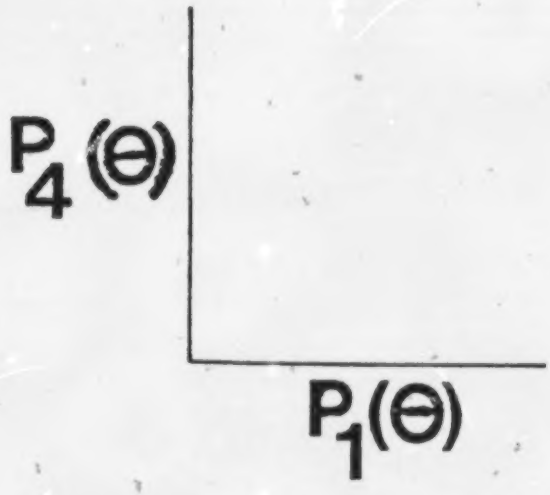




SAMPLE A

SAMPLE B

II. ITEMS 4 & 5 SIMILAR



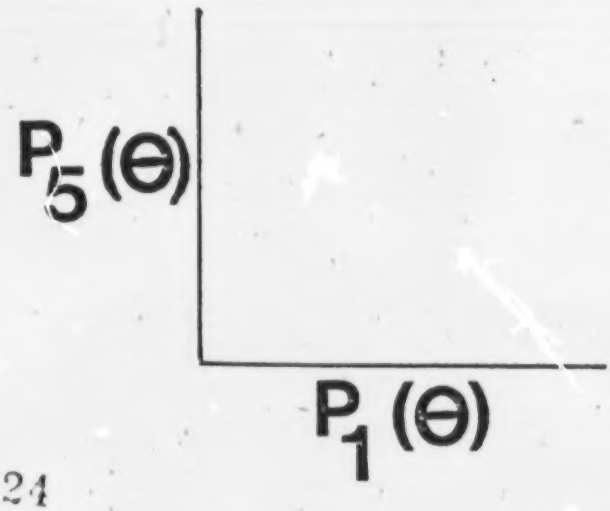
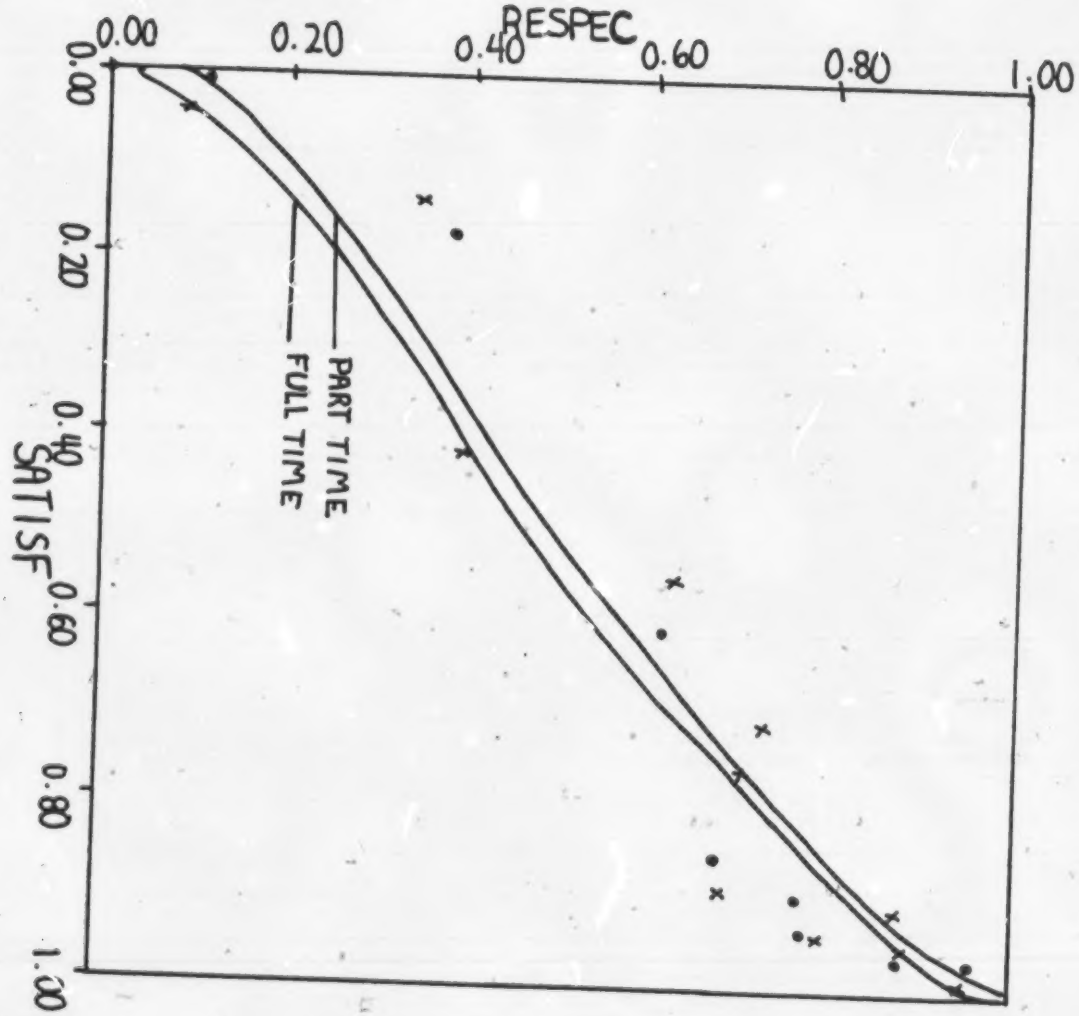


FIGURE 6



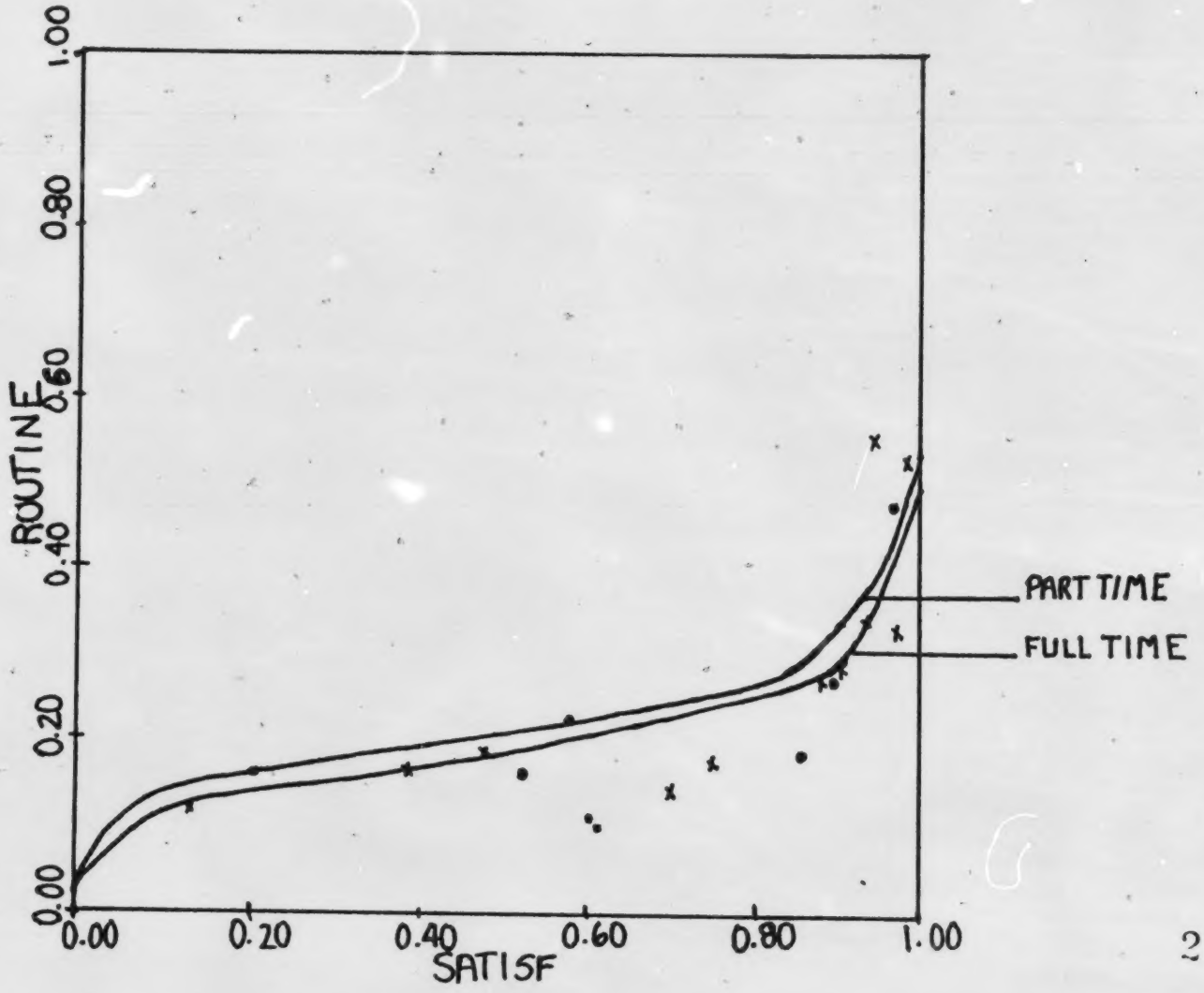

ROUTINE
1.00
0.80
0.60
0.40
0.20
0.00
0.00
0.20
0.40
0.60
0.80
1.00
SATISF
PART TIME
FULL TIME

FIGURE 7

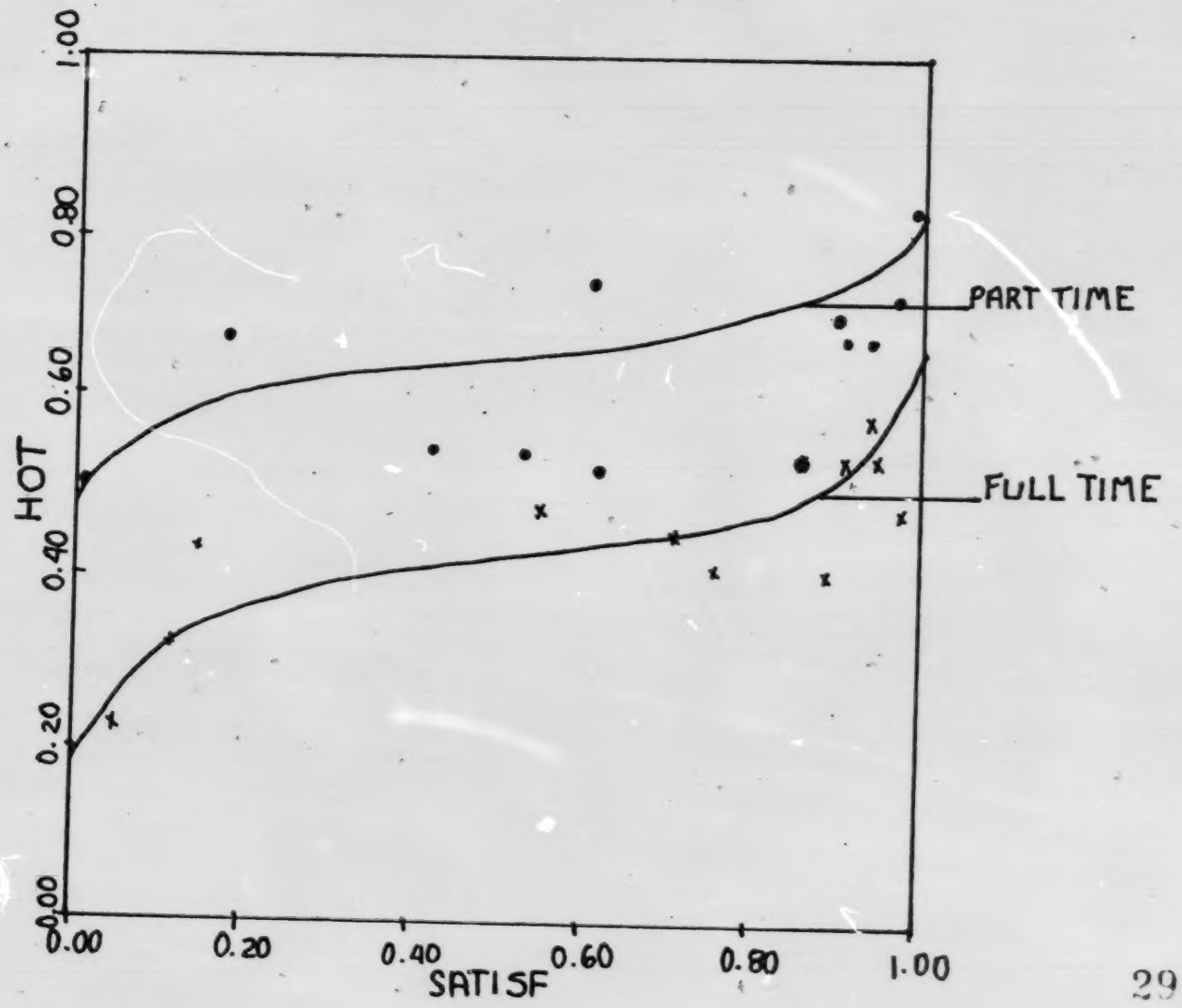

1.00
0.80
0.60
0.40
0.20
0.00
HOT
0.00
0.20
0.40
0.60
0.80
1.00
SATISF
PART TIME
FULL TIME

FIGURE 8

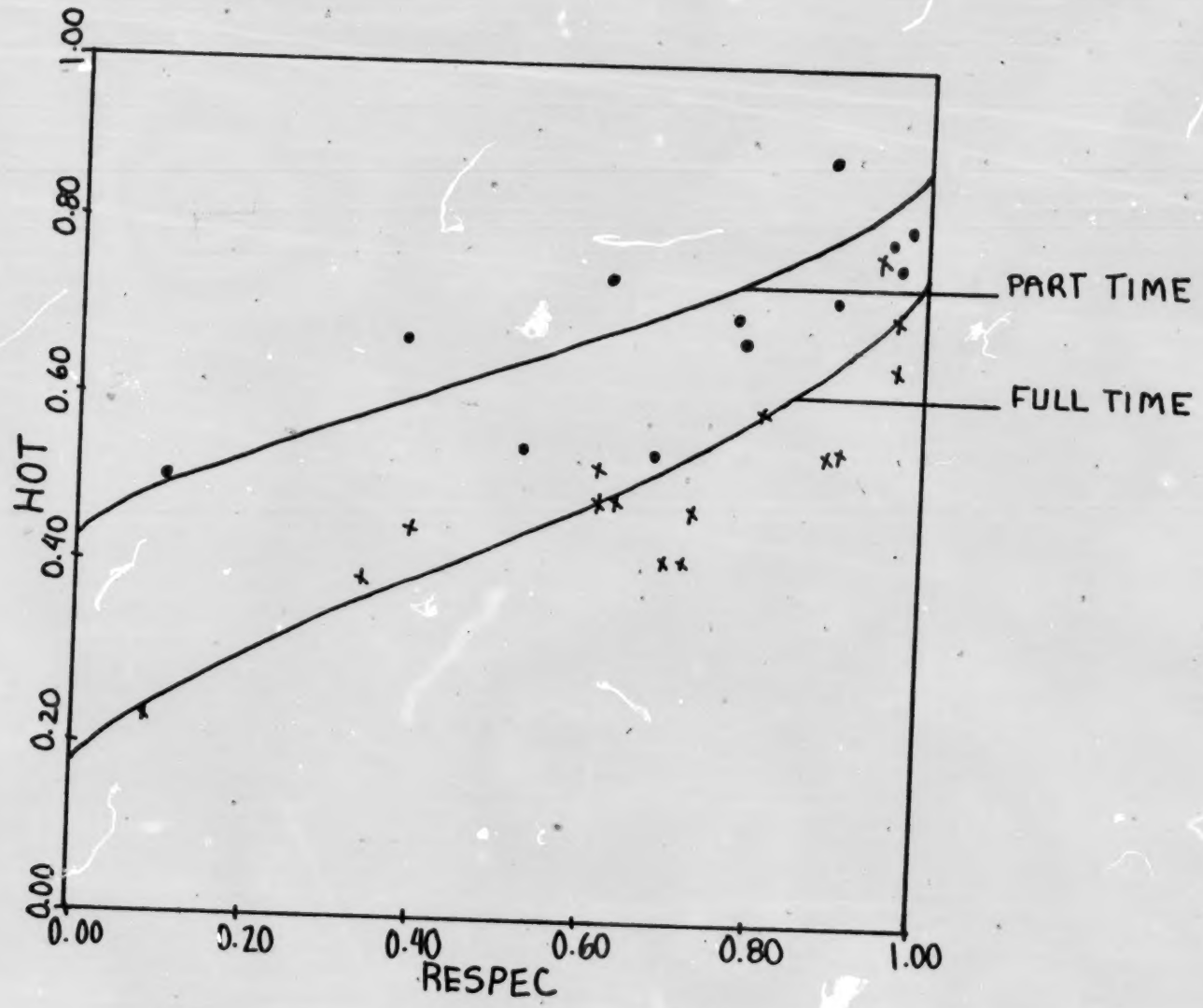

1.00
0.80
0.60
0.40
0.20
0.00
HOT
0.00
0.20
0.40
0.60
0.80
1.00
RESPEC
PART TIME
FULL TIME

FIGURE 9

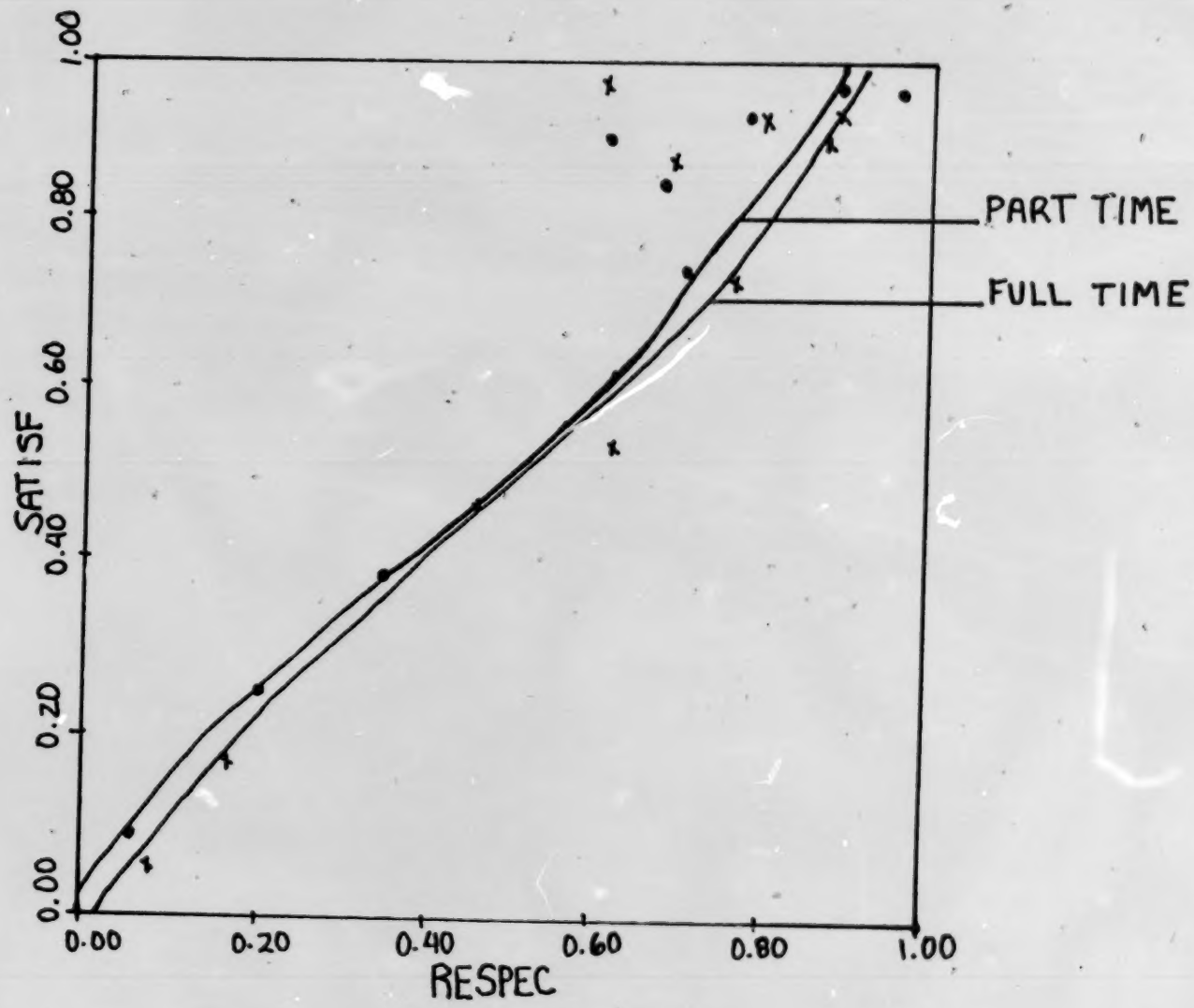

1.00
0.80
0.60
0.40
0.20
0.00
SATISF
0.00
0.20
0.40
0.60
0.80
1.00
RESPEC
PART TIME
FULL TIME

FIGURE 10

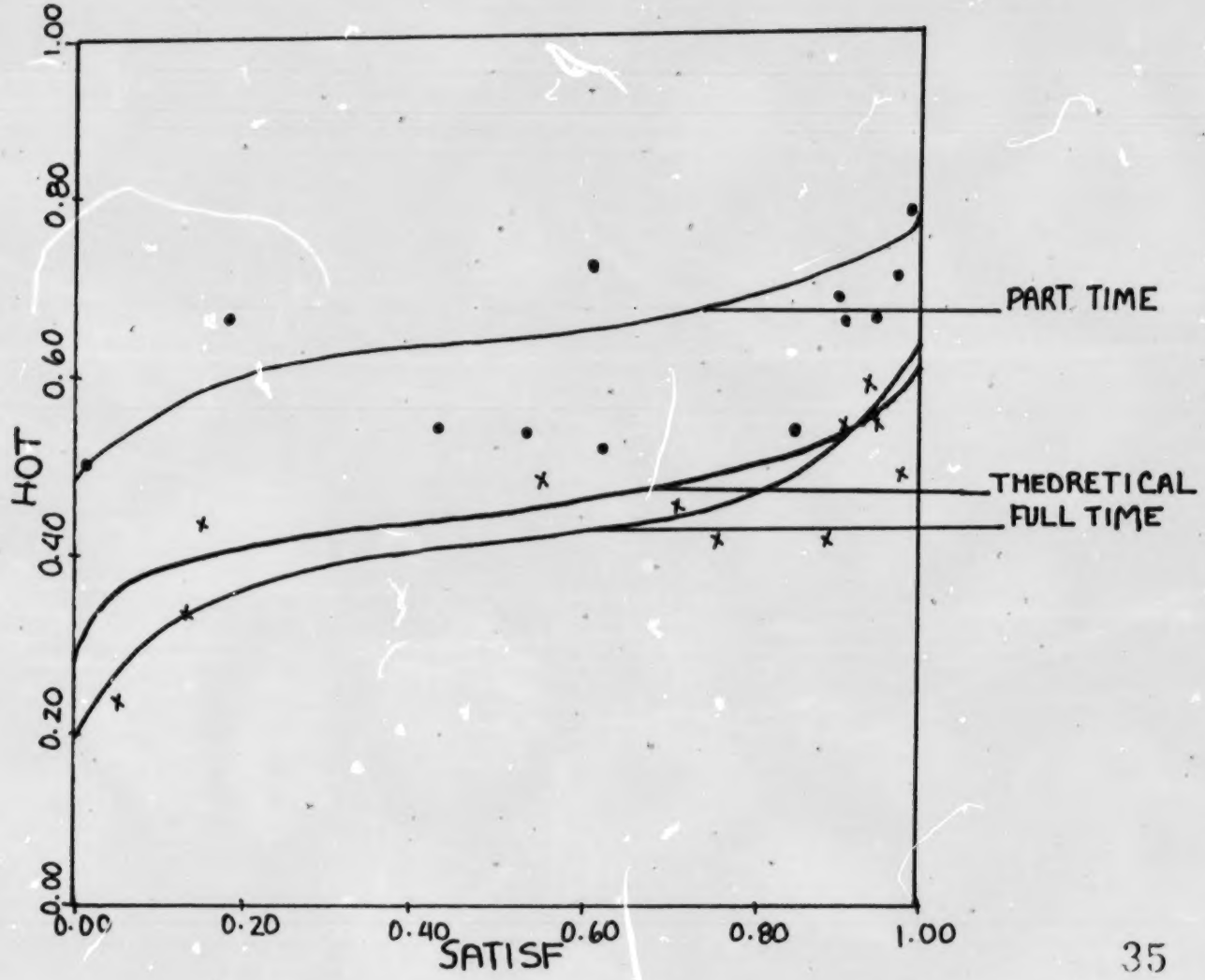

1.00
0.80
0.60
0.40
0.20
0.00
HOT
0.00
0.20
0.40
0.60
0.80
1.00
SATISF
PART TIME
THEDRETICAL
FULL TIME

FIGURE 11

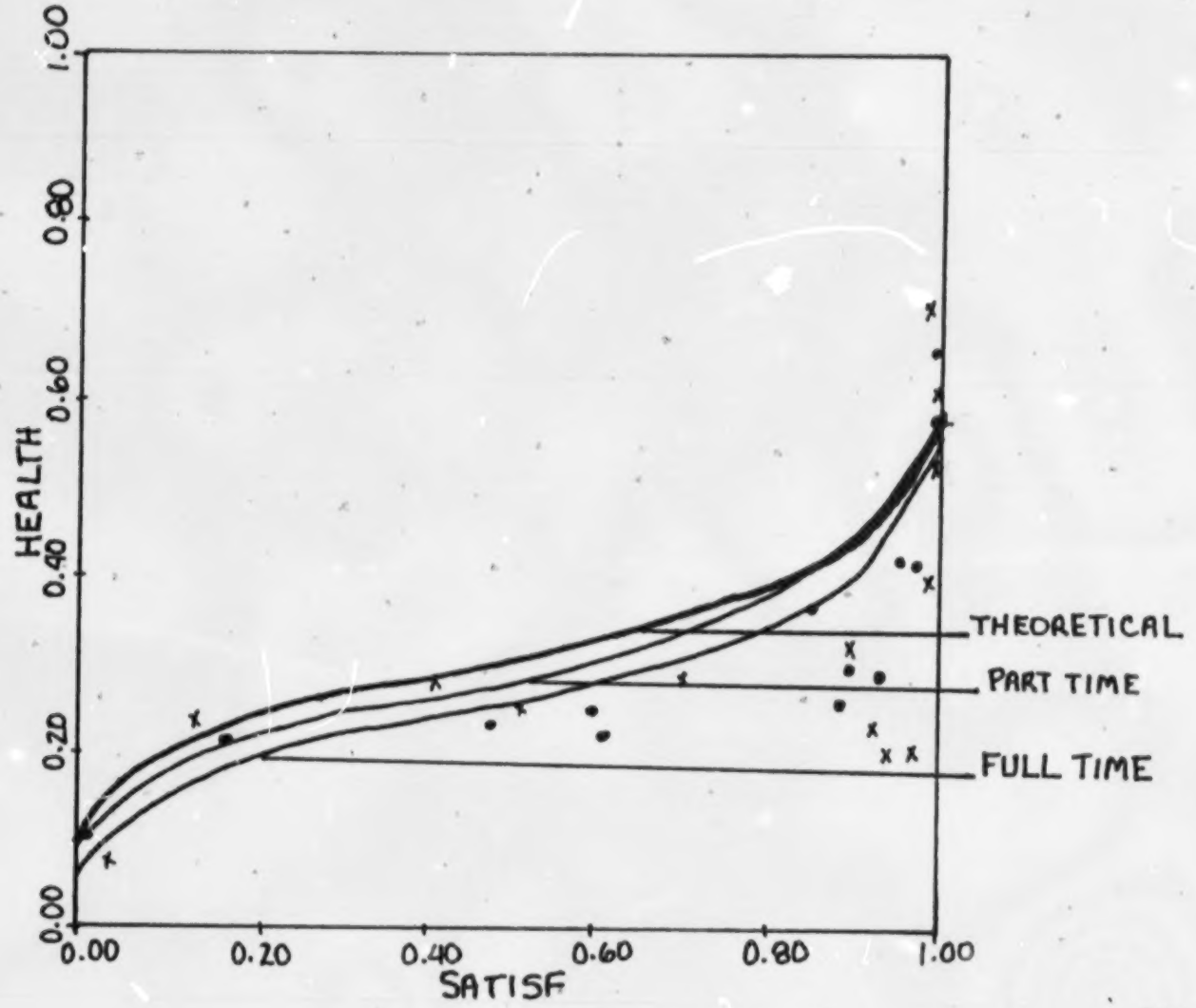

HEALTH
1.00
0.80
0.60
0.40
0.20
0.00
0.00
0.20
0.40
0.60
0.80
1.00
SATISF
THEORETICAL
PART TIME
FULL TIME

FIGURE 12

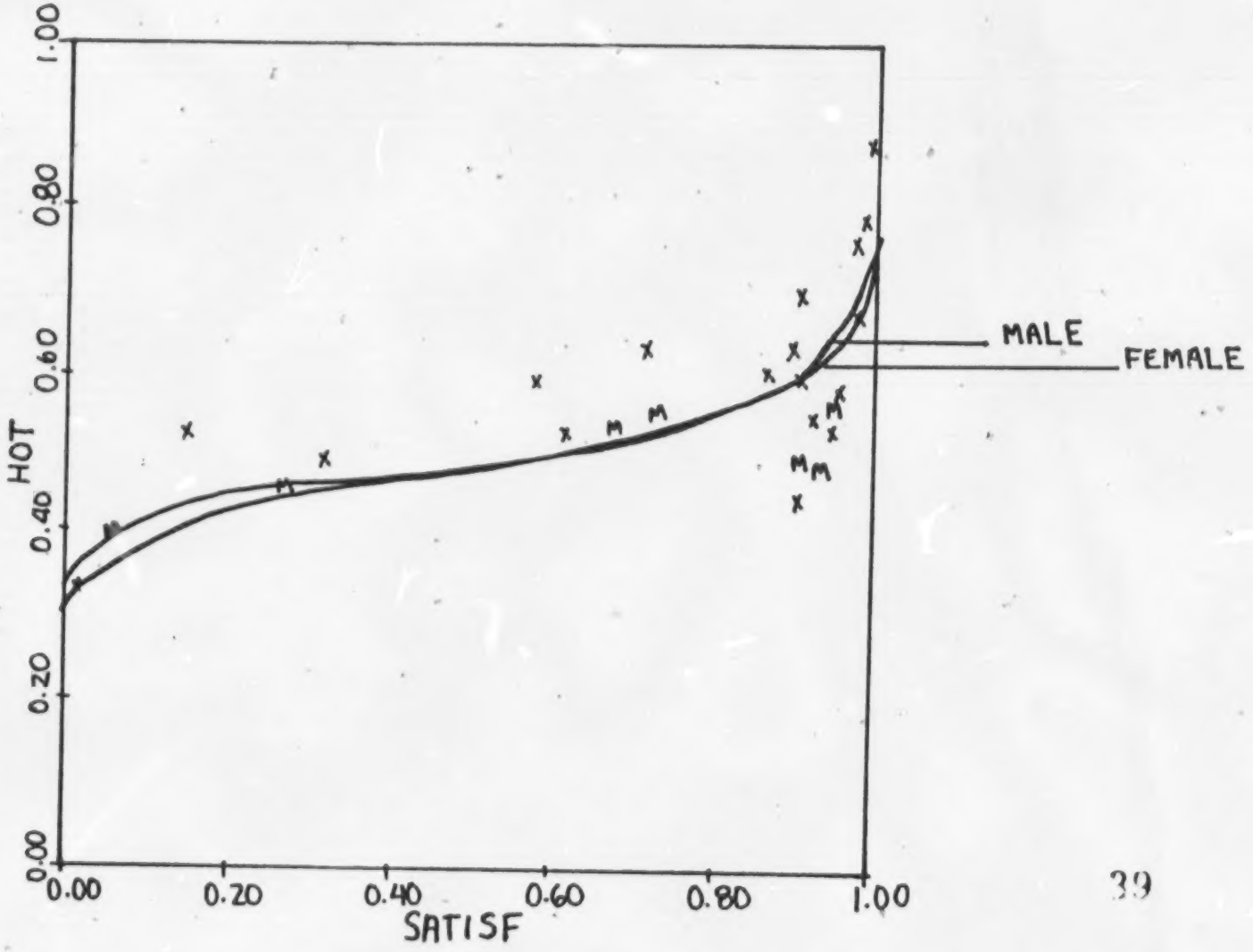
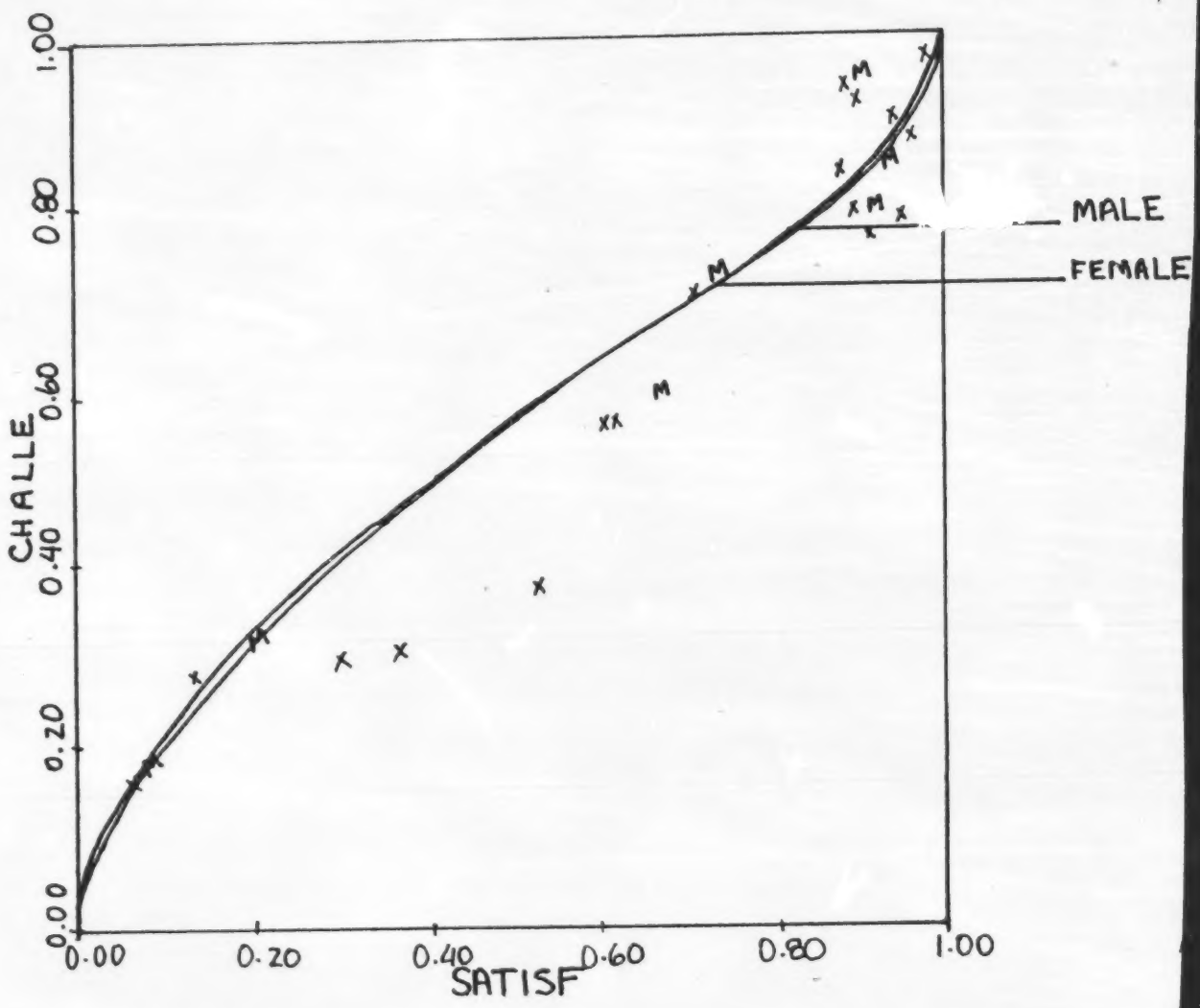

1.00
0.80
0.60
0.40
0.20
0.00
HOT
0.00
0.20
0.40
0.60
0.80
1.00
SATISF
MALE
FEMALE

FIGURE 13


CHALLE
SATISF
0.00
0.20
0.40
0.60
0.80
1.00
MALE
FEMALE